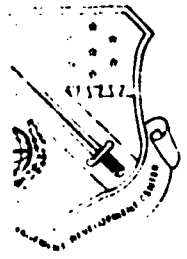
LMDC-TR-83-5

NOVEMBER 1983



AD A137340

AN EVALUATION OF ORGANIZATION DEVELOPMENT INTERVENTIONS: A LITERATURE REVIEW

DANIEL E. BOONE

NOVEMBER 1983

2004 0405 000

LEADERSHIP AND MANAGEMENT DEVELOPMENT CENTER
AIR UNIVERSITY
Maxwell Air Force Base, Alabama 36112

LMDC-TR-83-5

Technical Reports prepared by the Leadership and Management Development Center (LMDC), Maxwell Air Force Base, Alabama, report a completed research project documented by literature review references, abstract and testing of hypotheses, whether stated or implied. Technical Reports are intended primarily for use within the Air Force, but may be distributed to researchers outside the USAF, both military and civilian.

The views and opinions expressed in this document represent the personal views of the author only, and should not in any way be construed to reflect any endorsement or confirmation by the Department of Defense, the Department of the Air Force, or any other agency of the United States Government.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Public Affairs (PA) in accordance with AFR 190-17 and is releasable to the National Technical Information Center where it will be available to the general public, including foreign nations.

This Technical Report has been reviewed and is approved for publication.

LAWRENCE O. SHORT, Major, USAF
Chief, Research Operations

LLOYD WOODMAN, JR., Lt Col, USAF
Director, Research and Analysis

JOHN E. EMMONS
Colonel, USAF
Commander

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) LMDC-TR-83-5	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Leadership and Management Development Center (AU) 6b. OFFICE SYMBOL (If applicable) AN	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code) Maxwell Air Force Base, AL 36112	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Southeastern Center for Electrical Engineering Education 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F 49620-82-C-0035
8c. ADDRESS (City, State and ZIP Code) St. Cloud, Fla.	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT NO.
11 TITLE (Include Security Classification) An Evaluation of Organization Development: A Literature Review	
12. PERSONAL AUTHOR(S) Daniel E. Boone	
13a. TYPE OF REPORT Final 13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Yr., Mo., Day) November 1983 15. PAGE COUNT 51
16. SUPPLEMENTARY NOTATION	
17. COSATI CODES FIELD / GROUP / SUB. GR	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) organizational development (OD) OD evaluation, OD intervention, OD evaluation instrument

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This literature review covers those issues and elements necessary to implement a proper evaluation of the effects of an organization development (OD) intervention. Brief discussions of the purposes of OD interventions and evaluations are given, followed by a description of each step necessary in carrying out an evaluation. An optimal evaluation takes into account each of the following steps: the definition of goals; the selection of criteria; the selection of measures and problems associated with measuring change; threats to internal validity and research design; and statistical analysis of data collected. Each of these steps is discussed in detail, followed by an assessment of the current state of the art of OD research in light of these considerations. Recommendations are then made as to how OD evaluations can be improved in light of the discrepancies found between the "state of the art" and the ideal or optimal evaluation.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Captain Janice M. Hightower, USAF	22b. TELEPHONE NUMBER (Include Area Code) (205) 293-7034 22c. OFFICE SYMBOL LMDC/AN

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

TABLE OF CONTENTS

Introduction

Organization Development (OD) is a term used to describe a wide range of social-science based approaches to planned organizational change (Porras & Berg, 1978a). OD is a planned, systematic process of organizational change based on behavioral science technology, research, and theory (Beckhard, 1969; Hellreigel et al., 1973; Herrington, 1976). The practice of OD is aimed toward improving the quality of life for members of human systems and increasing the institutional effectiveness of those systems (Alderfer, 1977; Herrington, 1976). With the organization functioning below its capacity, it is the purpose of OD to determine the ultimate causes of these undesirable symptoms and then to devise ways to eliminate or at least minimize the cause(s) (Armenakis, Feild, & Holly, 1976). Eliminating the causes of undesirable symptoms, then, are the objectives of each OD intervention.

Organization Development can be defined as a sustained, long-range process of planned organizational change using reflexive, self-analytic methods of improving the functioning of an organizational system (Bennis, 1969; Campbell, Downas, Peterson, & Dunnette, 1974; Cook, 1976; Miles & Schmuck, 1971) with emphasis on improvement of an organization's problem solving and renewal processes with the assistance of a consultant or "change agent" (French & Bell, 1973). OD consultants help people do preventive maintenance on their relationships, problem solving abilities, and organizational structures, policies, and procedures (Weisbord, 1981).

OD places more emphasis than do other approaches (e.g., management development) on a collaborative process of data collection, diagnosis, and action for arriving at solutions to problems (Burke & Schmidt, 1970; Cook, 1976; Hellriegel, 1973; Herrington, 1976). Improvement of a dysfunctional organizational state implicitly involves standards or criteria for optimal performance. Even though what is considered to be optimal will differ from organization to organization, all OD efforts will be similar in attempting to identify these goals, objectives, and criteria for optimal performance. (Campbell et al., 1974; Cook, 1976).

This process has been labelled as "action research" (Campbell et al., 1974; French, 1982; Friedlander & Brown, 1974; Hellriegel, 1973; Nicholas, 1979; Weisbord, 1981) and underlies most of the interventions that have been invented in the evolution of OD (French, 1982; Hellriegel, 1973). The focus of action research has recently shifted from exploration, inquiry, and discovery to deliberate alteration and improvement of organizational structures through purposeful planning and systematic methodologies (Weisbord, 1981). The action research process involves problem identification, consultation, data gathering, diagnosis, feedback to the client, and data gathering after action (French, 1969).

The final evaluative stage collects data to monitor, measure, and determine effects which are fed back to clients for re-diagnosis and new action (Nicholas, 1979). To evaluate whether or not stated objectives and goals have been achieved, the presence or absence of causes and symptoms, i.e., the criteria, must be determined during the evaluation phase of the OD intervention (Armenakis et al., 1976). The design of OD research and the

measurement of effects from interventions can be viewed as a broader activity called evaluation research (Alderfer, 1977; Burke & Schmidt, 1970). The purpose of evaluation research is to measure the effects of a program or intervention against the goals the program set out to accomplish to improve future programming (Weiss, 1972). This means that the role of the researcher is to determine whether the changes in the system are the result of the OD effort or the result of extraneous occurrences. To justify the time and money expended in the OD intervention as well as to allow for the determination of the most effective technique of intervention (Franklin, 1976; Nicholas, 1979, Schuman, 1967), the researcher must attempt to establish cause and effect (De Meuse & Liebowitz, 1981) and to understand the underlying processes contributing to the observed effect.

Lewin (1946) emphasized the role of evaluation in action research as follows:

> If we cannot judge whether an action has led forward or backward, if we have no criteria for evaluating the relation between effort and achievement, there is nothing to prevent us from making the wrong conclusions and to encourage the wrong work habits. Realistic fact-finding and evaluation is a prerequisite for any learning (p.35).

Johnson (1970) and Hawkridge (1970) have distinguished between summative and formative evaluations. The primary purpose of a summative evaluation is to determine an overall evaluation of a program as it already exists. Formative evaluations use data collected during the development and initial tryout of a program as a basis for improving the program. Despite this distinction,

all evaluations more or less follow the procedural outline provided by Hawkridge (1970). According to Hawkridge, the seven phases of evaluation research are as follows: (1) setting up goals and objectives for the evaluation, (2) selecting objectives to be measured, (3) choosing instruments and procedures, (4) selecting samples for the intervention, (5) establishing measurement and observation samples, (6) choosing analysis techniques, and (7) drawing conclusions and recommendations. Each of these steps will be discussed in detail in the following sections. Issues that must be taken into consideration for each phase will be presented. Steps # 4 and # 5 will be considered together as part of the larger discussion on research design and methodology. Analysis techniques (#6) will also be covered in this design and methodology section.

The Process of Evaluation

Setting Up Goals and Objectives for the Evaluation

Evaluation of a program's effectiveness is not possible unless intended impacts of the program are stated in clearly measurable terms (Marqulies, Wright, & Scholl, 1977). In planning for the evaluation of an organization development intervention, it is important that specific goal selection be accomplished to assure that data appropriate to measurement of selected goals will be available or attainable (Hahn, 1970). A program objective or goal is simply an intended impact of the program itself on some target population. To specify an objective clearly, one must state the operations by which it can be determined whether and to what extent the objectives have been

obtained (Johnson, 1970). These operations are then the measures that are needed (Fitzpatrick, 1970; Carver, 1970). That is, if objectives are precisely (usually behaviorally) stated (Campbell et al., 1974; Carver, 1970; Franklin, 1976; Hahn, 1970; Johnson, 1970), the measurement problem is all but solved (Carver, 1970).

Selecting Criteria to be Measured

According to Porras and Patterson (1979, p. 41), "perhaps the most pressing issue in OD assessment research is the problem of which variable to measure." They further state that "it is easy to advocate that the assessors should measure the variables being affected by the intervention, but frequently we do not know what these variables are ahead of time." Many writers have emphasized the use of "hard" objective, behavioral measures (Armenakis & Feild, 1975; Armenakis et al., 1975), but frequently OD interventions also plan to impact or change attitudes also, sometimes referred to as "soft" criteria.

Armenakis and Feild (1975) further distinguish between internal and external hard criteria. The distinction lies in the degree to which the criteria are influenced by changes occurring external to the organization. Internal hard criteria are minimally influenced by changes occurring external to the organization and are readily accepted by organizational members as measures of organizational performance (e.g., productivity). External hard criteria would be considered influenced by these external changes. For

example, Georgapoulos and Tannenbaum (1957) have noted that "net profit... is a poor criterion in view of the many unanticipated fluctuations external to the system, e.g., fluctuations in the general economy, market sales, and earnings," (p. 535).

Campbell et al., (1974) have identified several dependent variables that could be assessed in the evaluation of organizational effectiveness. Although initially suggested as effectiveness criteria, they can also be applied as potential outcome criteria of a program intervention evaluation, depending on the objectives of the intervention (Fitzpatrick, 1970). Many of these can be classified as to whether they are a "soft" or "hard" criterion measures. A criterion will be considered "hard" if its measurement can potentially be obtained through objective, preferably behavioral, indices. Consensual agreement of attitudinal precepts will be considered to be "hard" within the context of this definition since the determination of consensual agreement should be relatively objective. "Soft" measures will be those involving subjective, attitudinal ratings for variables having no easily identifiable or observable criterion.

Hard criteria include:

Productivity. Productivity refers to the quantity or volume of the major product or service the organization provides.

Efficiency. This could be represented as a ratio that reflects a comparison of some aspect of unit performance to the costs incurred for that performance.

Profit. Profit is the amount of revenue from sales left over after all costs and obligations are met.

Accidents. This refers to the frequency of on the job accidents resulting in lost time.

Growth. Growth refers to the increase in such things as manpower, facilities, assets, and innovations.

Absenteeism.

Turnover. Turnover refers to any change of personnel within the organization.

Control. Control refers to the degree and distribution of management type of control that exists within an organization for influencing and directing the behavior of organization members.

Goal Consensus. This refers to the degree to which all individuals perceive the same goals for an organization.

Role and norm congruence Role and norm congruence refers to the degree to which the members of an organization are in planned agreement on such things as what kinds of supervisory attitudes are best, performance expectations, morale, role requirement, etc.

Managerial task skills. Refers to the overall level of skill the commanding officer, managers, or group leaders possess for performing tasks centered on work to be done, and not the skills employed when interacting with the organizational members.

Soft measures include:

Satisfaction. Satisfaction could be described as an individual's perception of the degree to which he or she has received an equitable amount of the outcome provided by the organization.

Morale. Morale is a predisposition in organizational members to put forth extra effort in achieving organizational goals and objectiveness.

Readiness. Readiness is an overall judgment concerning the probability that the organization could specifically perform some specified task if asked to do so.

Below are additional criteria listed by Campbell et al. (1974) that could entail both "soft" and "hard" measurement procedures:

Cohesion/Conflict. Cohesion refers to the extent that organization members like one another, work well together, communicate freely and openly, and coordinate their work efforts. Conflict refers to verbal and physical clashes, poor coordination, and ineffective communication.

Flexibility/Adaptation. This refers to the ability of an organization to change its standard operating procedures in response to environmental changes.

Managerial/Interpersonal Skills. Refers to the level of skills and efficiency with which management deals with supervisors, subordinates and peers and includes the extent to which management gives support, facilitates constructive interaction, and generates enthusiasm for meeting goals and achieving excellent performance.

Information management and communication. Refers to the collection, analysis, and distribution of information critical to organizational effectiveness.

Choosing Instruments and Procedures

Related to the question of which variables are to be measured is the question of how to go about measuring the variables (Gordon & Morse, 1975). Many of the variables that have been identified have several different operational forms. Existing records, direct observation, retrospective ratings by independent observers, and self-perceptions have all been used as sources for data (Franklin & Thrasher, 1976). Techniques include direct observation, the use of tests and questionnaires, the use of physical evidence data, and the use of archival records. According to Webb, Campbell, Schwartz, and Sechrest (1966), each of these techniques must be viewed in terms of its "obtrusiveness" or reactive effect on the subject or program participant. Obtrusiveness refers to the subject's awareness that he or she is being measured or observed, therefore affecting the behavior of interest. An obtrusive measure alters the natural course of the behavior as it would have occurred without the observation, i.e., a "guinea pig effect." Within this context, self-report questionnaire data and direct observation are the most obtrusive, while the use of archival measures and physical evidence are the least obtrusive. These last two techniques minimize the need to disturb subjects and lessen the extent to which the measurement process itself changes the behavior of interest.

Many have advocated the use of unobtrusive measurement techniques in the evaluation of OD interventions, but few have suggested specific procedures for carrying this recommendation out (Cummings, Molloy, & Glen, 1977). Direct observation is very costly and could be highly obtrusive, but does not rely

on a subject's retrospective account or his or her subjective impressions. Archival records are unobtrusive, cheap to obtain, easy to sample, and the population restrictions associated with them are often knowable However, Campbell (1969, p. 415) warns that "those who advance the use of archival measures as social indicators must face up not only to their high degree of chaotic error, but also to the politically motivated changes in record keeping that follow upon their public use as social indicators."

Although behavioral indices of change are preferred to the less objective measures, they are almost always more difficult and costly to obtain, hence the continual reliance on self-report questionnaire instruments. Questionnaires are relatively inexpensive and allow the collection of data from a large sample simultaneously. In addition, they easily lend themselves to statistical analysis, a feature lacking to a great degree with unobtrusive physical accretion and erosion techniques. However, Pate, Nielsen, and Bacon (1977) warn that "the exclusive use of questionnaire instruments in the assessment of organizational change capitalizes on chance outcomes and does not allow the researcher to obtain convergence on his or her results." They also add that "such practice does not permit the researcher to adequately handle the problem of response bias" (p. 454).

The use of questionnaires also brings up to the psychometric considerations of reliability and validity (Morrison, 1978). Reliability refers to the consistency with which a measuring device yields identical results when measuring identical phenomenon. Validity is concerned with how well a measure captures the essence of the phenomenon of interest. Issues of reliability and validity must be seriously considered whenever a "tailor-made"

questionnaire is developed according to the needs of a specific organization (Armenakis, Feild, and Holley, 1976). Often the demonstration of reliability and validity will take longer than the requirements of expediency of the OD intervention allow for. Gordon and Morse (1975) warn that "the lack of sensitive, validated, and reliable measurement instruments limits current attempts at evaluation" (p. 343). The requirements of a test instrument's reliability and validity will frequently compromise "tailor-made" measurement devices. Several other problems arise from the use of the questionnaire as the source and means of data collection (Alderfer, 1977; Carver, 1970; Golembiewski, Billingsley, and Yeager, 1976a; Pate et al., 1977). For example, Carver (1970) states that principles that have been validly developed for measuring between individual differences are invalidly used for measuring within individual change or group differences. His conclusion rests on the contention that items excluded in the construction of a measurement device to assess individual differences are the very items that should be included in order to determine whether any change took place. Due to the way the test was constructed, in tests of individual differences the relationship between the test score and the variable measured are not linear. Thus, a difference or change detected at one end of the scale may not reflect the same difference at another locale on the scale.

Golembiewski et al. (1976a) proposed that the entire concept of change is in need of clarification, particularly as it is accomplished through survey and questionnaire techniques. According to Randolph (1982, p. 119), "a unitary concept of change may be inappropriate and misleading, both in terms

of over-estimation and under-estimation of organizational change." Changes from OD interventions may involve any one or all of the following three conceptually distinct types of change as recently operationalized by Golembiewski et al. (1976a): Alpha, Beta, and Gamma change. Terborg et al. (1980, p. 111) state that "it is important to understand which type of change has occurred if the effects of interventions are to be unambiguously examined." Without an assessment of these change types, OD researchers may be led to conclude that a situation is deteriorating or that no change has occurred when in fact change has occurred (Alderfer, 1977; Armenakis, Feild, & Holley, 1976; Golembiewski, et al., 1976a; Lindell & Drexler, 1979; Macy & Peterson, 1983; Porras & Patterson, 1979; Randolph, 1982).

Gamma change occurs when the subject, over time and as a result of the OD intervention, changes his or her understanding of the criterion being measured (Zmud & Armenakis, 1978). This type of change involves a redefinition of concepts previously defined (Alderfer, 1977; Armenakis & Smith, 1978; Porras & Patterson, 1979). For example, if a factor analysis of a questionnaire indicates that several items measure a specific dimension, say leadership, then a factor analysis of a data set obtained subsequently should produce the same items measuring the same dimension, i.e., the factor structures should be identical. However, it could be that the planned OD intervention was directed or intended to enhance the subjects' understanding of the concept of leadership. If subjects have redefined a criterion during a change program, then questionnaire responses before the intervention may have little resemblance to responses after intervention and a comparison of responses would be meaningless and/or misleading (Armenakis & Smith, 1978; Armenakis & Zmud, 1979).

Beta changes are changes in perceptions of a dimension as determined by a measuring instrument in which scale intervals have varied over time. Beta change can occur when no actual behavior change is recorded as a change by respondents. Suppose that a supervisor at a second measurement is no more or less supportive than he or she was at the first measurement. One still might find a change in the supervisor's score on the scale if those who rated him or her changed the way they used the scale (Lindell & Drexler, 1979). In using self-report questionnaires, researchers assume that individuals using them in evaluating themselves or the situation have an internalized standard for judging their level of functioning with regard to a given dimension, and that this internalized standard will not change from pretest to posttest. Researchers must be able to state what each particular score on the pretest set of scores is equivalent to on the posttest set of scores, i.e., a common metric must exist between the two sets of scores (Cronbach & Furby, 1970). If the standard of measurement changes between the pretest and posttest, the two ratings will reflect this difference in addition to changes attributable to the experimental manipulation (Howard, Schmeck, & Bray, 1979). Consequently, comparisons of the ratings will be invalid. This threat to the internal validity of evaluation design has also been referred to as "instrumentation" by Campbell and Stanley (1966) and as "the response shift bias" by Howard and Dailey (1979).

Alpha change is that change which is detected along a consistent measurement scale (i.e., no beta change) and for which gamma change has been ruled out (Alderfer, 1977; Armenakis & Zmud, 1979; Golembiewski & Billingsley, 1980; Lindell & Drexler, 1979; Porras & Patterson, 1979; Zmud & Armenakis,

1978). In other words, the phenomenon itself, and neither the subject's understanding of it nor the scale units has changed (Armenakis & Zmud, 1979). Alpha change takes place when an actual behavioral change is recorded as such by respondents. For example, a change occurs when a respondent, on a leadership scale, indicates leader behavior as changing from a "2" to a "3" when in fact the leader's behavior has changed by that amount (Armenakis & Smith, 1978).

These prior explanations and illustrations allow for a fuller understanding of the formal definitions of alpha, beta, and gamma change as initially provided by Golembiewski et al. (1976a) and referred to by many others (Golembiewski & Billingsley, 1980; Lindell & Drexler, 1979; Macy & Peterson, 1983; Roberts & Porras, 1982; Porras & Patterson, 1979):

> ALPHA CHANGE involves a variation in the level of some existential state, given a constantly calibrated measuring instrument related to a constant conceptual domain.
>
> BETA CHANGE involves a variation in the level of some existential state, complicated by the fact that some intervals of the measurement continuum associated with a conceptual domain have been recalibrated.
>
> GAMMA CHANGE involves a redefinition or reconceptualization of some domain, a major change in the perspective or frame of reference within which phenomena are perceived and classified, in what is taken to be relevant in some slice of reality (p. 134).

Differentiating alpha, beta, and gamma change is of special importance to researchers because this typology is closely intertwined with the objectives of behavioral interventions (Alderfer, 1977; Armenakis & Zmud, 1979; Zmud

& Armenakis, 1978). If the purpose is to improve leader behavior and to reflect this improvement by measuring subordinate perceptions of leader behavior, then alpha change may be intended (Armenakis & Zmud, 1979). On the other hand, the purpose might be to change respondents' understanding of leadership and then gamma change may be intended. Golembiewski and Billingsley (1980) state that "gamma change constitutes the goal of many planned interventions" because OD seeks to change "the concepts of the quality of organization life that should and can exist" (Golembiewski et al., 1976a).

Alpha, beta, and gamma change may be caused by the sources of invalidity and/or the OD effort (Armenakis & Smith, 1978). A true experimental research design would help to determine whether the OD intervention caused the observed changes (experiments and research designs will be discussed more fully later). However, comparison group designs are often impossible in organization development research. It is the absence of a comparison group in combination with the use of questionnaires that can result in a difficult determination of the presence and degree of change that occurred as a result of the organization development intervention. Many authors have addressed this issue (Armenakis & Zmud, 1979; Golembiewski & Billingsley, 1980; Macy & Peterson, 1983; Randolph, 1982; Terborg, Howard, & Maxwell, 1980), suggesting a two step process in order to determine the effects of an OD intervention as a result of alpha change: 1) detect gamma change first, for if it exists, beta and alpha change cannot be detected; 2) if it can be shown

that gamma change has not occurred, beta change must be then assessed, for if it exists alpha change cannot be assessed. Only if gamma and beta change are discounted can alpha change be assessed.

Golembiewski et al. (1976b) suggest testing for differences in the factorial structures of measures across time as an operational way to determine whether gamma change took place. Zmud and Armenakis (1978) describe the rationale (Ahmavaara, 1954) for using the procedure as follows:

> Since gamma change involves a redefinition of criterion being investigated, subject response structures (as determined through factor analysis) that result from each administration of the measurement device must be compared (p. 666).

This comparison would entail the amount of common variance shared between the pre- versus post-intervention structures (Golembiewski & Billingsley, 1980). A very high congruence between before and after structures signals that no gamma change has occurred. Golembiewski and Billingsley (1980) have set a cutoff of 50 percent common variance or less as indicating the possibility of gamma change occurring, while Macy and Peterson (1983) state that if the common variance is greater than 85 percent, it can be safely concluded that any measured changes are not gamma changes.

Zmud and Armenakis (1978) have offered a methodology for assessing beta change using questionnaires. They suggested that alpha and beta changes can be differentiated when pre and post ratings are collected both on actual and ideal criterion levels. Through comparison of actual scores, ideal scores, and

differences between actual and ideal scores, they maintain it is possible to infer alpha or beta change, assuming no gamma change. If ideal scores have changed, respondents have recalibrated the measurement scale (Randolph, 1982). Examination of difference scores will clarify whether beta changes or both alpha and beta changes have occurred.

If gamma and beta change are discounted, the next step is to assess for alpha change. Terborg, Howard, and Maxwell (1980) suggest that this be done by using t-test comparisons of mean differences between treatment and comparison groups.

Lindell and Drexler (1979, p. 14) maintain that the importance of Golembiewski's conceptual distinctions of change are "substantially overstated." Their argument lies in asserting that changes in factor structure can also be attributed to alpha and beta changes (therefore demonstrating the insignificance of gamma change considerations), and that beta change will not occur if "psychometrically sound" instruments are used, i.e., tests consisting of reliable scales consisting of multiple items with behavioral anchors. Having dispensed with gamma and beta change, Lindell and Drexler (1979, p. 18) argue that consideration of alpha change alone is sufficient, "since there is little doubt that a psychometrically sound questionnaire needs to be interpreted as anything other than face value."

In response to Lindell and Drexler's first point, Golembiewski and Billingsley (1980, p. 101) state that "our critics have obviously missed the point" regarding alpha change since "by our definition, alpha change implies

no appreciable change between pre and post intervention factorial structures." In support of Golembiewski and Billingsley's argument, Randolph (1982) has since demonstrated that gamma change can occur without alpha change. In response to Lindell and Drexler's second point, Golembiewski and Billingsley (1980) state their critics fail to recognize that the present state of OD assessment technology does not meet the "psychometrically sound" criteria, nor do they provide for a means of detecting beta change, given the likelihood of its occurrence.

The time at which the post intervention measurement is taken is an issue that also must be considered (Armenakis, Feild, & Holly, 1976). Measurements taken immediately after an intervention may reflect more clearly specific learnings from the program. On the other hand, delayed measuring may show that effects which initially appeared to be strong have weakened or disappeared. Porras (1977) found that the longer the time between the end of the active intervention process and the last measurement of the research variables the fewer significant changes were reported. Morrison (1978, p. 43) states that "practitioners hold that OD is an ongoing process and not a time-bound intervention, and therefore traditional means of evaluation do not apply."

In summary, "the measurement process needs much innovation and development" (Porras & Patterson, 1979, p. 56). The measurement process stands at the interface between the respondent's behaviors and attitudes and the researcher's abstraction of those phenomena. Porras and Patterson (1979) state that despite its critical role in this linkage process, "our abilities to measure adequately are not receiving heavy emphasis or concentrated development" (p. 56).

In the meantime, the use of multiple measures is advocated, providing convergent evidence that an accurate assessment is being made concerning the presence or absence of the variable of interest (Campbell et al., 1974; Cummings et al., 1977; Fitzpatrick, 1970; Golembiewski et al., 1976; Pate et al., 1977; Webb et al., 1966). As will be discussed more fully later, the majority of OD interventions rely on self-report questionnaires in the data collection phase of the evaluation. Pate et al. (1977) state that "the exclusive use of questionnaire instruments in the assessment of organizational change capitalizes on chance outcomes and does not enable the researcher to obtain convergence of his or her results" (p. 457). Webb et al. (1966) state that:

> The mistaken belief in the operational definition of theoretical terms has permitted social scientists a complacent and self-defeating dependence upon single classes of measurement, usually the interview or questionnaire. Yet the operational implication of the inevitable theoretical complexity of every measure is exactly opposite: it calls for multiple operationalism, that is, for multiple measures which are hypothesized to share in the theoretically relevant components but have different patterns of irrelevant components.

The advantage of using more than one mode of measurement is the opportunity to determine the method variance in the measurement, thus providing a more accurate determination of the variable's true value, and hopefully more insight about the variable itself (Campbell et al., 1974). From this perspective, the

use of self-report attitude questionnaires is relevant, provided that they are used as a sample of the total measurement universe (Fitzpatrick, 1970).

Research Design and Methodology

As stated earlier, the purpose of evaluation research is to measure the effects of a program or intervention against the goals the intervention set out to accomplish. The researcher must determine the presence of change as well as establish a causal connection between the program intervention and the subsequent effects. Plans for carrying these tasks out are referred to as research designs. The following section will discuss the strengths and weaknesses of some of the research designs frequently used by OD evaluators.

The ultimate test of the strength of any research design relates to its internal and external validity. (Armenakis et al., 1976; Campbell, 1969; Campbell & Stanley, 1966; Cummings et al., 1977; Duncan, 1981; Evans, 1975; Morrison, 1978; Posavac & Carey, 1980; Staw, 1980). The research design is internally valid if it allows the researcher to eliminate alternative explanations or rival hypotheses relative to the intervention and the outcome. Campbell (1969) states that the mere possibility of some alternative explanation is not enough - it is only the plausible rival hypotheses that are invalidating. If one can confidently state that the intervention program caused the observed effects the design is internally valid. If the results obtained can be accurately generalized to other subjects, situations, and settings, the design is externally valid.

External validity asks whether the experiment's findings can be generalized beyond the specific population, environment, and operational definitions of the independent and dependent variables used in the study (Cummings et al., 1977). Campbell and Stanley (1966) have identified four threats to external validity:

Interaction effects of testing. This refers to the effects of a pretest in modifying a subject's responsiveness to the program intervention, thus threatening any generalization to an unpretested population.

Interaction effects of selection and treatment. The treated population may be more responsive and hence unrepresentative of the universal population.

Reactive effects of the experimental arrangements. This refers to the artificiality of the experimental setting which makes it atypical of settings to which the treatment is to be regularly applied.

Multiple treatment interference. This refers to the interaction between several different programs taking place simultaneously.

Threats to Internal Validity

Campbell and Stanley (1966) have also identified nine threats to internal validity, also referred to as sources of invalidity. Since threats to internal validity are the primary concern of program evaluators (Posavac & Carey, 1980), future discussion of research designs and their attempts to establish the effect of a planned organization development intervention will focus exclusively on these internal threats. These threats can be understood as possible research errors that can make the determination of cause and effect difficult if not impossible (Duncan, 1981). The nine threats as identified by Campbell and Stanley (1966) are as follows:

History. History refers to those events, in addition to the program intervention, which occur simultaneously between the first and second measures in the dependent variable and thus provide an alternative explanation for the changes observed. It is a change that affects the organizational unit but is not related to the OD effort. Using some type of comparison group that does not receive the program but is exposed to the same historical events should control for this threat to internal validity.

Maturation. Maturation refers to changes within an organization as a unit and/or its members as a function of the organization's or individual's own natural development, that are independent of the OD effort and are operating as a function of the passage of time. Many authors point out the lack of using long-term follow-up procedures in assessing the presence or absence of an effect resulting from an OD intervention. Some conclude that, given the state of the art of present OD evaluation methodology, maturation as a possible source of internal validity is of little concern since the evaluation assessments cover only a short period of time. If the individual is taken as the unit of analysis, adults (versus children) employed by an organization would be expected to have already attained a steady state of maturity. Howeve this assertion can only be applied to physical maturation, and maturation as ar, potential source of invalidity must not be so casually eliminated from consideration. Campbell's (1963) "continued improvement" thesis demonstrates how development maturation can occur at the organizational level. The thesis states that any reliable organization is expected to improve its performance naturally, by virtue of the organization's purpose to achieve a common goal (Armenakis & Feild, 1975).

Instability. Instability refers to the unreliability of a measure. This can apply to questionnaires that are necessarily imperfect measures of the criterion and to human judges who may use inconsistent standards or grow fatigued with an increasing number of observations of the criterion behavior.

Testing. Testing is defined simply as the effects of taking a test on the scores of a subsequent test. Taking a pretest could sensitize a respondent and subsequently influence his or her responses on the posttest. An example of this would be the Hawthorne effect in which the subjects react to obtrusive measurement techniques. The early Hawthorne studies demonstrate that when intact work groups are singled out for special attention, changes in the dependent variable may not be wholly attributable to changes in the independent variable (White & Mitchell, 1976). Another example of testing would be when participants in the OD effort attempt to respond to subsequent administrations of the same or similar questionnaires differently because the first administration made them sensitive to what was desired by the change agent. Margulies et al. (1977) state that if the pretest is influential in focusing attention to problem areas, it should be considered as part of the planned OD intervention. A way to control for the effects of testing would be to use nonreactive, unobtrusive measures (Webb et al., 1966) and to collect "hard" data (Golembiewski et al., 1976).

Instrumentation. Instrumentation refers to changes in the calibration of a measuring instrument or changes in the observer which result in changes in the obtained measurements. As an example, the "new broom" that introduces abrupt changes of policy is also apt to reform the record keeping procedures,

and thus confound reform effects with instrument change (Campbell, 1969). As another example, an organizational member exposed to a program designed to improve team cohesiveness may indicate that his organization has changed from a "2" to a "3" on a cohesiveness scale when in fact no change has occurred. In each of these examples, the standard of measurement has changed, i.e., the instrument has been recalibrated.

Statistical regression. This threat to internal validity refers to the movement of an individual's extreme score toward the mean on a subsequent administration of the assessment device. In the field of organization development, subjects or groups are often selected for participation in the intervention program because of a state of need as reflected in their extreme scores on an assessment device (Campbell, 1969). Organizations seeking services from OD consultation programs are often severely deficient in a desired area or lacking in areas of standard performance. Groups scoring poorly on the first administration of a test are likely to have as one component of their low score an extreme error term that depresses their score. On a subsequent administration of the test, the extreme conditions contributing to the poor score are not likely to be present to the same degree that they were at the initial administration. The absence of these depressing factors will enhance the score; hence, the score regresses toward the overall mean (the reverse logic applies in the instance of an extremely high score). A change that is due to statistical regression may be confused with a change produced by the intervention.

Selection. Selection refers to biases resulting from differential recruitment of comparison groups, producing different mean levels on the measures of the effects. This source of internal invalidity often occurs with the nonrandom assignment of subjects to treatment and comparison groups. Any changes in the effectiveness of the organization could be explained by the initial differences in relevant characteristics in the two groups. As a result of their initial differences, the two groups may have differed on the outcome criterion measure regardless of the OD intervention.

Experimental mortality. Mortality refers to the differential loss of respondents from the groups being observed.

Interaction effects. This threat to internal validity refers to the instance when two or more of the above errors interact to confound the results of a research design. The interaction effect most commonly referred to as a threat to the internal validity of an experimental design is the selection-maturation interaction, where differential rates of maturation or autonomous change occur as a result of selection bias.

In addition to the threats enumerated by Campbell & Stanley (1966), another group of possible confounds exist that will be referred to as "relationship effects." These effects refer to the relationship (usually unconscious) between consultant and participant that serves to enhance the likelihood of a positive outcome. For example, if organizational members know and respect the change agent, a halo effect could occur causing subjects to supply the desired results regardless of the intervention employed. The potential for halo effects provides strong argument for the use of external consultants and evaluators, i.e., those individuals not directly involved

with the organization receiving the treatment. Other possible relationship effects include placebo, Pygmalion, and experimenter demand effects.

Considerable attention has been devoted to the development of research designs which can be applied to assess whether an observed change can be attributed to the OD intervention (Campbell & Stanley, 1966; Cook & Campbell, 1979). These research designs can be evaluated in terms of the degree to which they control for the various sources of internal invalidity (Howard et al., 1979; Margulies, 1977). The more precisely a research design controls for these errors the more adequate it becomes. In a totally artificial laboratory situation most of the errors can at least be measured and their influence on the criterion of interest can be considered. However, as the laboratory environment is removed, problems begin to develop in controlling the sources of invalidity. The further removed from the laboratory, the less rigorous the evaluation methodology becomes. Terpstra (1981) has found that the number of positive evaluations of organization development interventions increases as the methodological rigor of the designs decrease. Bass (1983) suggests that research outcomes in the less rigorous designs can just as easily be attributed to investigator bias, and to placebo, Hawthorne, and Pygmalion effects on the participants. Gordon and Morse (1975) have similarly concluded that the more rigorous designs are less likely to produce positive results. In conclusion, positive results obtained from the more rigorous research designs are more likely to indicate true intervention effects. Therefore, program evaluators must attempt to employ more rigorous research methodology (Armenakis et al., 1975; Cummings et al., 1977; Macy & Peterson, 1983; Margulies et al., 1977; Pate et al., 1977; Porras & Berg,

1978(a); Porras & Patterson, 1979; Randolph, 1982; Terpstra, 1981; White & Mitchell, 1976).

What follows is a discussion of the research methodologies or experimental designs available to the program evaluator. First there will be a presentation of the ideal situation (the "true" experiment) followed by a discussion of some of the research designs that are open to many rival hypotheses. Finally, a discussion of some compromising designs, i.e., those designs that are not "true" experiments but do control for many of the threats to internal validity, will be presented.

Research Designs

True experiment design True experimental designs are thought to control for all threats to internal validity (Bentler & Woodward, 1979; Campbell & Stanley, 1966; Franklin, 1977; Staw, 1980). However, Cook and Campbell (1979) state that experimentation does not control for these threats to internal validity: imitation, compensatory equalization, and compensatory rivalry. These threats occur because of the difficulty in truly separating a control and an experimental or treatment group in an organizational setting. For example, if a supervisor finds out about an intervention occurring in another work group, he or she may behave differently in order to compensate, thus preventing a true control.

In the classic experimental design (also called the pretest - posttest control group design) one first establishes the independent and dependent variables of interest and decides how they are to be measured or varied. Subjects are then chosen randomly from some larger and defined population and

assigned by random means to two (or more) subgroups. Different "treatments" representing different aspects of one or more of the independent variables are then applied to the various groups while one or more groups remain "untreated;" i.e., they serve as controls for the experimental procedure. The effect of the exposure to the program is determined by comparing any changes in those exposed to the treatment with changes in those not exposed (Weiss, 1972). Campbell et al. (1974) states that "judiciously timed measurement of the dependent variable across the several groups and analysis of differences among the measurements yield inferences about the causal effects of different levels of the independent variable on the dependent variable" (p. 174).

The strengths of this design are achieved by randomization and the use of a control group. Randomization prevents systematic differences in the initial status of the experimental and control groups. A substitute procedure commonly used when randomization is not feasible is matching subjects on relevant characteristics. However, Weiss (1972) states that program evaluators are often unable to define the characteristics on which people should be matched. Cook and Campbell (1979) also warn that matching as a substitute for randomization can result in regression effects. For example, a group that is lacking in some desirable characteristic might seek an intervention. A pretest is given in order to determine the group's level on this desirable characteristic. The comparison group in most cases will not be lacking in the characteristic of interest to the same degree that the treatment group is (if the comparison group was deficient to a similar degree it too would have sought the intervention). Thus, individual scores from the comparison group should be relatively higher than individual scores from the

treatment group. If the program evaluator decides to match subjects on the basis of their pretest scores, the matched subjects will represent different ends of the distribution of their respective group. A relatively low score in the comparison group would be matched with a relatively high score from the treatment group. Due to statistical regression, the low scores from the comparison group will regress toward the mean of the comparison group on a subsequent posttest. In a similar fashion, the high scores from the treatment group will regress toward the mean of the treatment group on a subsequent posttest. Two scores that were once equal are now drastically different on the b s of statistical regression alone. This difference or change is commonly mistaken for evidence that an OD intervention was effective.

Complete randomization of subjects to groups provides the tremendous advantage of assuming that the groups so assigned do not significantly differ from one another prior to the intervention (Fuqua, 1979). Thus by randomly assigning subjects to experimental and control groups, any differences between these two groups observed after the experimental group has been exposed to an intervention which were not observed during the pretest can be attributed to the effects of the intervention. In fact, in a truly randomized situation, there is no necessity to show that the groups were equivalent through the use of a pretest (Campbell & Stanley, 1966). Assuming that randomization to groups insures similarity, many have argued against the use of the pretest (Campbell, 1957; Linn & Slinde, 1977). These authors state that often the act of an initial observation itself (as in a pretest) is reactive (i.e., a source of internal invalidity defined previously as testing). Adding pretests to the design also weakens its validity because the pretest may have interacted

with the actual program to cause the observed change. This gives rise to the posttest only control group design, to be used if randomness is assured.

Methods of analyzing experiments include: 1) a t-test to test the significance between the difference between mean scores for the treatment and control groups; 2) a simple ANOVA to simultaneously compare the means of three or four groups to learn whether at least one of them is different from the other means; 3) complex ANOVA to study the effects of more than one factor simultaneously; 4) if a pretest is given, an ANCOVA could be used using the pretest score as the covariate.

Although true experimentation ranks highest in terms of providing valid causal inference, it is not always the most practical course of action in organizational settings. Only in rare instances are evaluators able to exercise the amount of control required for experimental designs (Franklin, 1976). The experimental design is exceedingly difficult to apply in actual field settings because of the experimental requirements of randomization and control group use and the many other unplanned events and interventions occurring differentially across groups (Campbell et al., 1974; Evans, 1975). Individuals cannot always be assigned to experimental and control groups because such a procedure might disrupt normal population systems or produce inequities between experimental and control groups and hence be considered unethical.

Thus, difficulties associated with true experimental designs limit their usefulness in organization development evaluations. Given the practical problems inherent in experimental designs, a number of alternatives to the

experimental design have been suggested (Campbell & Stanley, 1966; Cook & Campbell, 1979).

A simple and commonly used design is the one group pretest/posttest design. In this design, observations are made before and after an intervention is introduced to a single group. This design can indicate whether any change has taken place, but is not rigorous enough to allow the assessment of the intervention's causal connection to the observed changes. The design is open to many potential rival hypotheses, including history, maturation, testing, instrumentation, mortality, and statistical regression.

Quasi-experimental designs. Because of the limitations of the one group pretest/posttest designs and the impracticality of true experiments quasi-experimental designs are frequently employed (Campbell & Stanley, 1966; Duncan, 1981; Friedlander & Brown, 1974). According to Weiss (1972), quasi-experimental designs have the overriding feature of feasibility and can produce results that are sufficiently convincing of an intervention's causal connection with observed changes. Unlike true experiments designed to rule out the effects of influences other than exposure to the program, quasi-experimental designs often depend on the possibility that these influences can be ruled out by statistical techniques (Linn & Slinde, 1977). Instead of randomly assigning subjects to groups, quasi-experimental designs utilize intact group that are likely to be different or "nonequivalent" on many variables.

One of the most popular quasi-experimental designs is the time series experiment (Armenakis et al., 1976; Armenakis & Smith, 1978; Campbell &

Stanley, 1966; Cook & Campbell, 1979, Franklin, 1976; Weiss, 1972). The essence of the time series design is the presence of a periodic measurement process on a single group that acts as its own control both prior to and after the introduction of an intervention. The effect of the intervention is indicated by a discontinuity in the measurements recorded in the time series. This design does not account for the potential confound of history, i.e., some other event besides the intervention could account for the observed discontinuity. History could be controlled if a comparison group is employed. Maturation is more or less controlled for if the time series is extended. It is not likely for a maturation change to occur between measurements in the time series after the intervention that did not occur before the intervention. In a similar way instrumentation can be accounted for. Selection and mortality are ruled out if the same specific persons are involved at all observations. Regression effects are usually a negatively accelerated function of elapsed time (Campbell, 1969) and are therefore implausible as explanations of an effect after the intervention that is greater than the effects between pretest observations.

As many pretest and posttest measures of the evaluation criteria should be made as possible. Simple comparisons of one or two pretest scores with one or two posttest scores may be influenced by extremes and therefore be misleading. Armenakis and Smith (1978) recognize that the use of many measurements is necessary in order to eliminate with confidence many of the threats to internal validity and to assess the immediate and extended impact

of the intervention. However, the usefulness of this design is limited if repeated measures are made using the questionnaire approach solely. The effects of testing as a source of internal invalidity would be compounded as the number of repeated observations are made. Respondents may be sensitized to the nature of the changes to be expected if the same assessment device is repeatedly used. If, to decrease this possibility, the between observation time intervals are lengthened, ruling out the effects of history become even more difficult (Franklin, 1976). In order to avoid the problems associated with reactivity to a series of questionnaire measurements, Macy and Peterson (1983) argue for the use of archival and behavioral data in the manner outlined by Webb et al. (1966). Armenakis and Smith (1978) and Terborg et al. (1980) advocate the use of a reduced number of observations.

Statistics for assessing change in time series designs must account for the fact that data collected in an organizational setting is often not independent; i.e., adjacent measures in the series have a higher correlation than non-adjacent points (Armenakis & Feild, 1975). This phenomenon is referred to as autocorrelation (Campbell, 1963; Macy & Peterson, 1983: Tryon, 1982) and has been discussed by Cronbach and Furby (1970).

Armenakis and Feild (1975) use a regression technique to determine the significance of the difference between pretest and posttest measurements. A trend line is calculated for the data before the intervention, after the intervention, and one for the entire research period. Variances from these trend lines are then calculated and an F ratio is produced. From these, it

is determined if there were any significant statistical differences between pretest and posttest performance.

Tryon (1982) uses the C statistic to determine whether the time series contains any trends, i.e., systematic departures from random variation. The logic underlying the C statistic is the same as the logic underlying visual analysis; variability in successive data points is evaluated relative to changes in slope from one phase of the time series to another. The C statistic aids the evaluator in evaluating how large the squared deviations from the mean are (which reflect the presence of all types of trends) relative to the sum of the squared consecutive differences (which are independent of all types of trends). The logic of this fraction is analogous to that of the F statistic.

Another common quasi-experimental design is the nonequivalent comparison group design (Campbell & Stanley, 1966; Evans, 1975; Franklin, 1976; Fuqua, 1979). This design is similar to the one group pretest/posttest design but is different in that it employs a comparison group. The term "nonequivalent" arises from the fact that subjects are not randomly assigned to the program or comparison groups. Instead, in this design groups represent intact units. Consequently, this design presents the potential for treatment and comparison groups which differ significantly from one another before the intervention. The more similar the intervention and comparison groups are in their recruitment, and the more this similarity is confirmed by pretest scores, the more effective this design is in controlling the sources of invalidity.

Including comparison groups permits a distinction to be made between the effects of the program and the several alternate plausible interpretations of change. Both treatment and comparison groups will have had the same amount of time to mature, historical events will have affected both equally, testing effects would be the same since both groups were tested twice, and mortality could be examined equally for both groups. The main problem of the nonequivalent control group design is not selecting a comparison group sufficiently similar to the intervention group. For example, people choosing to enter a program are likely to be different from those who do not, and the prior differences might make post-intervention comparisons tenuous. Nonequivalent control group designs are especially sensitive to regression effects when the treatment group has been selected on the basis of an extreme score on a pretest (Evans, 1975).

The problem presented by the nonequivalent control group design basically consists of eliminating group differences which exist at pre-intervention assessment from the analysis of group differences at post-intervention assessment. Reichardt (1979) provides a concise review of the literature which proposes analytic techniques for use with nonequivalent control group designs. The literature indicates that analysis of covariance (ANCOVA) procedures have received the most attention. ANCOVA is a statistical procedure for eliminating the effects of extraneous sources of variance from dependent measures, properly used only when it is not possible to use experimental controls to achieve the same result.

Although ANCOVA provides an attractive method for analyzing data from the nonequivalent control group design, it has not proved wholly adequate when

used for this purpose. For example, it has been demonstrated that measuring error can have a biasing effect on the analysis (Campbell & Erlebacher, 1970) and that under some conditions the analysis may either underadjust or overadjust for selection differences (Cronbach, Rogosa, Price, & Folden, 1976). At present there is no single method for analyzing data from the nonequivalent control group design that will be free of bias in all situations (Fuqua, 1979). Given the current state of analytic technology, Reichardt (1979) suggests that multiple analytic techniques be employed.

The combination of the time series design and the nonequivalent control group design yields a design that is more rigorous than either one by itself. This combination design had been given various names: the multiple time series design, the control series design, the modified time series design. The combination design is similar to the time series design but is different in that a comparison group is used. This added feature provides for a design that rules out all of the threats to internal validity (Campbell & Stanley, 1966; Franklin, 1976). The multiple measurements before the program is implemented will point out any differences existing between the two groups, facilitating the interpretation of any effects from the intervention. A variant of this design is the interrupted time series with switching replications (Cook & Campbell, 1979). This again is a time series design with a comparison group. In this design the comparison group receives the same intervention as the treatment group but at a later time.

Porras and Wilkins (1980) used a pooled regression approach to test for statistical differences between treatment and comparison group measures taken

at multiple points in time. Treatment and comparison group data was pooled to calculate the coefficients for one overall regression line. A "dummy variable" was then added which permitted the slopes of the two groups to be different. A third line was then estimated which permitted both slopes and intercepts to be different. The amount of variance explained by each of these resultant lines was represented by an R^2 for each regression. By comparing the R^2 for each of the three equations, a test was made to determine if letting the slopes or intercepts be different would give a better fit and thus explain a greater proportion of variance. If the third equation explained more of the variance than the first (pooled) one, then it could be concluded that the treatment group was performing differently than the comparison group. If the third equation showed a higher R^2 than the second, then the intercepts were different. If the second equation had a higher R^2 than the first, the slopes were different. Differences in intercepts and slopes were indicative of changes in behavior over time.

The State of the Art of OD Intervention Evaluations

Having described the essential ingredients necessary for proper evaluations, the remainder of this literature review will address whether and to what extent organization development research has addressed these issues. Porras and Berg (1978) state that relatively little OD evaluation research has been done, while others describe the current state of the art of OD evaluations as underdeveloped (Friedlander & Brown, 1974; Margulies et al.,

1977; Morrison, 1978; Pate et al., 1977). Part of the reason for this is that the field of organization development is a relatively new application of the behavioral sciences. Research methodology best suited for this new field is still in its developmental stages.

Administrative and methodological considerations also contribute to the systematic avoidance of proper OD intervention evaluations. On the administrative level, a divergence exists between the research-theory perspective of an evaluator and the action-change perspective of management. Management personnel involved in the planning and implementation of OD interventions are primarily concerned with answers to immediate problems (Pate et al., 1977). The pragmatic emphasis of "getting something useful" from the OD effort often places research in a secondary priority. Others hesitate to implement evaluation of existing programs for fear that evaluation process would interfere or change the process of development and change already taking place (Margulies et al., 1977).

On the methodological level, resistance arises out of the pessimism that exists in trying to implement a rigorous research design able to control for all sources of internal invalidity (Morrison, 1978). For example, organizational realities prevent random assignment to control and treatment groups (Armenakis et al., 1976; Macy & Peterson, 1983; Porras & Patterson, 1979). Field conditions of organization development research also prevent the full control of such extraneous variables and influences as the varying degrees of the intervention's implementation, multiple interventions taking place at once, and the time when post-intervention criterion measurements are taken. OD research also suffers from what is described as a "criterion deficiency

problem" (Armenakis et al., 1976; Porras & Berg, 1978b). Most standardized instruments used in OD evaluation research were not designed specifically to measure those variables and criteria that are frequently targeted in OD interventions. Thus, OD evaluators face a deficiency in the availability of measures for the criteria of interest.

Despite the difficulties mentioned above, OD evaluations have been attempted. Recent literature reviews by Armenakis et al. (1975), Cummings et al. (1977), De Meuse and Liebowitz (1981), Pate et al. (1977), Porras and Berg (1978a), Terpstra (1982), and White and Mitchell (1976) frequently come up with similar results but will arrive at markedly different conclusions. For example, Porras and Berg (1978a) state that there exists a reasonably large number of "scientific" investigations of the effects of OD programs. Two years later Porras and Wilkins (1980), using the same studies reviewed by Porras and Berg (1978a), conclude that there has been a slow rate of development of OD assessment and research methods.

In addition to arriving at different conclusions in the face of similar results, authors also reach different conclusions based on different results. For example, White and Mitchell (1976) conclude that most OD research uses poor research design while Porras and Berg (1978b) state that there is a large number of OD studies using research designs possessing a high degree of scientific rigor. A review of the OD literature thus does not consistently provide an unequivocal assessment of the "state of the art" of OD intervention evaluations.

A possible explanation for these diverse conclusions might be found in what could be described as "the floating criterion" phenomenon. Across the

various literature reiews, authors include studies only if they meet certain, predetermined selection criteria. These selection criteria vary from author to author, depending on a particular author's personal interpretation of what an OD evaluation should consist of. Thus, the criteria for selection (and the subsequent results and conclusions) "floats" or varies from author to author.

As an example of the occurrence of the "floating criterion," Porras and Berg's (1978a) selection criteria will be examined. Included in their sample of OD intervention evaluations were only those studies which: 1) used "human-processual" interventions; 2) were done in "representative segments" of real-life organizations; 3) measured organizationally relevant process variables; and 4) used quantitative techniques. Out of 160 evaluations surveyed, only 35 met these criteria. It is on these 35 studies that Porras and Berg (1978a) base their conclusion that the current OD evaluations are adequately rigorous in their research methodology. Had the remaining 125 studies been included, conclusions more similar to White and Mitchell's (1976) may have been reached.

In order for a clearer picture of the current status of OD evaluations to be drawn there must be agreed upon standards of what a proper OD evaluation consists of. The procedural outline provided by Hawkridge (1970) and used in this paper suggests the necessary ingredients of the proper evaluation. The literature reviews previously done by Armenakis et al. (1975), Cummings et al. (1977), De Meuse and Liebowitz (1981), Pate et al. (1977), Porras and Berg (1978a), Terpstra (1982), and White and Mitchell (1976) will be examined below from the perspective outlined by Hawkridge (1970), i.e., the type of criteria used, the type of measures used, and the type of research design used will be examined.

Criteria Measured

Most authors (Cummings et al., 1977; De Meuse & Liebowitz, 1981; Pate et al., 1977; Porras & Berg, 1978a; Terpstra, 1982; White & Mitchell, 1976) have found a predominant if not exclusive use of soft measures. Relatively few if any of the research employed made use of hard measures. An exception to this trend is Armenakis et al.'s (1975) finding that nearly three-fourths of the research surveyed used hard criterion measures. This finding is not surprising since the evaluations selected for this particular review were chosen from organizations that were primarily profit-oriented.

Instruments and Procedures Used

The use of subjective, attitudinal questionnaires is also common in current OD evaluation research (Armenakis et al., 1975; Cummings et al., 1977; De Meuse & Liebowitz, 1981; Pate et al., 1977; Porras & Berg, 1978a; Terpstra, 1982; White & Mitchell, 1976). Armenakis et al. (1975) and Porras and Berg (1978a) found that tailor-made questionnaires are frequently used when available. De Meuse and Liebowitz (1981), Porras and Berg (1978a), and Terpstra (1982) all found an absence in the use of longitudinal, follow-up measures.

Research Design and Methodology

White and Mitchell (1976) found that three-fourths of the research they reviewed failed to use components necessary in order to establish cause and

effect (e.g., comparison groups). Armenakis et al. (1975) also found an infrequent use of comparison groups. However, 44% of the studies reviewed by Armenakis et al. did employ the time series design, which is thought to control for many of the threats to internal validity (Campbell & Stanley, 1966). White and Mitchell seem to dismiss the time series design as a possible means of establishing cause and effect. Of those studies employing a comparison group, Armenakis et al. (1975) and Porras and Berg (1978a) found a frequent use of the modified time series design when a nonequivalent control group could be identified.

Recommendations

The literature reviews referred to above demonstrate some of the discrepancies that exist between what might be considered the optimal evaluation and the current state of the art of OD intervention evaluations. In light of these discrepancies the following recommendations are warranted:

1) More adequate instruments for measuring change related to OD intervention criteria should be developed.

2) In addition to "soft" criterion measures, "hard" criterion measures should be used.

3) In order to determine the long-term effects and maintenance of a program, multiple and longitudinal measurement should be carried out.

4) Where the selection of experimental and control groups on a random basis is not possible, the use of a comparison group-even an unmatched or nonequivalent group - should be used.

References

Adams, J., & Sherwood, J.J. An evaluation of organizational effectiveness: An appraisal of how Army internal consultants use survey feedback in a military setting. Group and Organization Studies, 1979, 4, 170-182.

Ahmavaara, Y. Transformation analysis of factorial data. Annals of the Academy of Science Fennicae (Series B), 1954, 881, 54-59.

Alderfer, C. P. Organizational development. Annual Review of Psychology, 1977, 28, 197-233.

Armenakis, A. A., & Feild, H. S. Evaluation of organizational change using nonindependent criterion measures. Personnel Psychology, 1975, 28, 39-44

Armenakis, A. A., Feild, H. S., & Holly, W. H. Guidelines for overcoming empirically identified evaluation problems of organizational development change agents. Human Relations, 1976, 29, 1146-1161.

Armenakis, A. A., Feild, H. S., & Mosley, D. C. Evaluation guidelines for the OD practitioner. Personnel Journal, 1975, 54, 99-103, 106.

Armenakis, A. A., & Smith, L. A practical alternative to comparison group designs in OD evaluation: The abbreviated time series design. Academy of Management Review, 1978, 3, 499-506.

Armenakis, A. A., & Zmud, R. W. Interpreting the measurement of change in organizational research. Personnel Psychology, 1979, 32, 709-723.

Bass, B. M. Issues involved in relations between methodological rigor and reported outcomes in evaluations of OD. Journal of Applied Psychology, 1983, 68, 197-199.

Beckhard, R. Organization development: Strategies and methods. Reading, MA: Addison-Wesley Publishing Co., 1969.

Bennis, W. G. Organization development: Its nature, origins, and prospects. Reading, MA: Addison-Wesley Publishing Co., 1969.

Bentler, P. M., & Woodward, J. A. Nonexperimental evaluation research: Contributions of causal modeling. In L. Datta and R. Perloff (Eds.), Improving evaluations. Beverly Hills, CA: Sage Publications, 1979.

Boje, D. M., Fedlor, D. B., & Rowland, K. M. Myth making: A qualitative step in OD interventions. Journal of Applied Behavior Science, 1982, 18, 17-28.

Burke, W. W. Organization development in transition. Journal of Applied Behavior Science, 1976, 12, 22-43.

Burke, W. W., & Schmidt, W. H. Primary target for change: Manager or organization. Organizational frontiers and human values. Belmont, CA: Wadsworth Publications, 1970.

Campbell, D. T. Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 1957, 54, 297-312.

Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), Problems in Measuring Change. Madison, WI: University of Wisconsin Press, 1963.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Campbell, J. P., Bownas, D. A., Peterson, N. G., & Dunnette, M. D. The measurement of organizational effectiveness: A review of relevant research and opinion. Minneapolis, MN: Personnel Decisions, Inc., 1974.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory Education: A national debate, the disadvantaged child, Vol 3. New York: Brunner Mazel, 1970.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluation research: Strategies and methods. Pittsburgh: American Institutes for Research, 1970.

Cook, J. H. Organizational development: An available management strategy. Carlisle Barracks, PA: U. S. Army War College, 1976.

Cook, T. D., & Campbell, D. T. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally, 1979.

Cook, T. D., & Reichardt, C. S. Statistical analysis of non-equivalent control group designs. Evaluation, 1976, 3, 136-138.

Cronbach, L. J., & Furby, L. How should we measure change: Or should we? Psychological Bulletin, 1970, 74, 68-80.

Cronbach, L. J., Rogosa, D., Price, G., & Folden, R. Analysis of covariance: Temptress and deluder or angel of salvation? Occasional paper, Stanford Evaluation Consortium, Stanford University, 1976.

Cummings, T. G., Molloy, E. S., & Glen, R. A methodological critique of fifty-eight selected work experiments. Human Relations, 1977, 30, 675-708

De Meuse, K. P., & Liebowitz, S. J. An empirical analysis of team-building research. Group and Organization Studies, 1981, 6, 357-378.

Duncan, W. J. Quasi-experimental research in organizations: A critique and proposed typology. Human Relations, 1981, 34, 989-1000.

Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-402.

Evans, M. E. Opportunistic organizational research: The role of patch-up designs. Academy of Management Journal, 1975, 18, 99-108.

Fitzpatrick, R. The selection of measures for evaluating programs. In Evaluation research: Strategies and methods. Pittsburgh; PA: American Institutes for Research, 1970.

Franklin, J. L., & Thrasher, J. H. An introduction to program evaluation. New York: John Wiley and Sons, 1976.

French, W. L. Organizational development: Objectives, assumptions, and strategies. California Management Review, 1969, 12, 23-24.

French, W. L. The emergence and early history of organization development. Group and Organization Studies, 1982, 7, 261-278.

French, W. L., & Bell, C. H., Jr. Organization development: Behavioral science interventions for organization improvement. Englewood Cliffs, NJ: Prentice Hall, 1973.

Friedlander, F., & Brown, D. L. Organization development. Annual Review of Psychology, 1974, 25, 313-342.

Fuqua, D. R. Measurement issues and design alternatives: Selecting criteria for the OD consultant. Improving Human Performance Quarterly, 1979, 8, 277-290.

Georgopoulos, M. S., & Tannenbaum, A. S. A study of organizational effectiveness. American Sociological Review, 1957, 22, 534-540.

Golembiewski, R. T., & Billingsley, K. R. Measuring change in OD panel designs: A response to critics. Academy of Management Review, 1980, 5, 97-103.

Golembiewski, R. T., Billingsley, K. R., & Yeager, S. Measuring change and persistence in human affairs: Types of change generated by OD designs. Journal of Applied Behavioral Science, 1976, 12, 133-157.(a)

Golembiewski, R. T., Billingsley, K. R., & Yeager, S. The congruence of factor analytic structures: Comparisons of four procedures and their solutions. Academy of Management Review, 1976, 1, 27-35.(b)

Gordon, G. & Morse, E. V. Evaluation research. Annual Review of Sociology, 1975, 1, 339-361.

Hahn, C. P. Methods for evaluating counter-measure intervention programs. In Evaluation research: Strategies and methods. Pittsburgh, PA: American Institutes for Research, 1970.

Hawkridge, D. G. Designs for evaluative studies. In Evaluation research: Strategies and methods. Pittsburgh, PA: American Institutes for Research, 1970.

Hellriegel, D., Slocum, J. W., Woodman, R. W. Organizational behavior (3rd ed.). St. Paul, MN: West Publishing Co., 1973.

Herrington, D. J. Air Force improvement through organization development (Technical Paper No. 1265-76). Maxwell Air Force Base, AL: Air Command and Staff College, 1976.

Howard, G. S., & Dailey, P. R. Response-shift bias: A source of contamination of self-report measures. Journal of Applied Psychology, 1979, 64, 144-150.

Howard, G. S., Schmeck, R. R., & Bray, J. H. Internal invalidity in studies employing self-report instruments: A suggested remedy. Journal of Educational Measurement, 1979, 16, 129-135.

Johnson, G. H. The purpose of evaluation and the role of the evaluator. In Evaluation research: Strategies and methods. Pittsburgh, PA: American Institutes for Research, 1970.

Kimberly, J. R., & Nielsen, W. R. Organization development and change in organizational performance. Administrative Science Quarterly, 1975, 20, 191-206.

King, A. S. Expectation effects in organizational change. Administrative Science Quarterly, 1974, 19, 221-230.

Lewin, K. Action research and minority problems. Journal of Social Issues, 1946, 2, 34-46.

Lindell, M. K., & Drexler, J. A. Issues in using survey methods for measuring organizational change. Academy of Management Review, 1979, 4, 13-19.

Linn, R. L., & Slinde, J. A. The determination of significance of change between pre- and post testing periods. Review of Educational Research, 1977, 47, 121-150.

Macy, B. A., & Peterson, M. F. Evaluating attitudinal change in a longitudinal quality of work life intervention. In S. E. Seashore, E. E. Lawler P. H. Nurvis, and C. Cammann (Eds.), Assessing organizational change: A guide to methods, measures, and practices. New York: John Wiley and Sons, 1983.

Margulies, N., Wright, P. L., & Scholl, R. W. Organization development techniques: Their impact on change. Group and Organization Studies, 1977, 2, 428-448.

Miles, M. B., & Schmuck, R. A. Improving schools through organization development: An overview. In R. A. Schmuck and M. B. Miles (Eds.), Organization development in schools. Pan Alto, CA: National Press Books, 1971.

Morrison, P. Evaluation in organization development: A review and an assessment. Group and Organization Studies, 1978, 3, 42-68.

Nicholas, J. M. Evaluation research in organization change interventions: Considerations and some suggestions. Journal of Applied Behavioral Science 1979, 15, 23-40.

Nunnally, J. The study of change in evaluation research: Principles concerning measurement, experimental design, and analysis. In E. L. Struening and M. Guttentag (Eds.), Handbook of evaluation research (Vol.1) Beverly Hills, CA: Sage Publications, 1975.

Pate, L. E., Nielsen, W. R., & Bacon, P. C. Advances in research on organizational development: Toward a beginning. Group and Organization Studies, 1977, 2, 449-460.

Porras, J. I. The comparative impact of different organization development techniques. (Research Paper No. 386). Stanford University: Graduate School of Business, 1977.

Porras, J. I., & Berg, P. O. Evaluation methodology in OD: An analysis and critique. Journal of Applied Behavioral Science, 1978, 14, 151-173.(a)

Porras, J. I., & Berg, P. O. The impact of organization development. Academy of Management Review, 1978, 3, 249-266.(b)

Porras, J. I., & Patterson, K. Assessing planned change. Group and Organization Studies, 1979, 4, 39-58.

Porras, J. I., & Wilkins, A. Organization development in a large system: An empirical assessment. Journal of Applied Behavioral Science, 1980, 16, 506-534.

Posavac, E. & Carey, R. G. Program evaluation: Methods and case studies. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1980.

Randolph, W. A. Cross-lagged correlational analysis in dynamic settings. Journal of Applied Psychology, 1981, 66, 431-436.

Randolph, W. A. Planned organizational change and its measurement. Personnel Psychology, 1982, 35, 117-139.

Reichardt, C. S. The statistical analysis of data from nonequivalent control group designs. In T. D. Cook and D. T. Campbell, Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally, 1979.

Roberts, N. C., & Porras, J. I. Progress in organization development research. Group and Organization Studies, 1982, 7, 91-116.

Rosenthal, R., & Rosnow, R. L. The volunteer subject. New York: John Wiley and Sons, 1975.

Sakamoto, G. H. A guide for assessing organizational change (Technical Paper No. 2095-79). Maxwell Air Force Base, AL: Air Command and Staff College, 1979.

Staw, B. M. The experimenting organization: Strategies and issues in improving causal inference within administrative settings. In E. E. Lawler, D. A. Nadler, and C. Cammann (Eds.), Organizational assessment: Perspectives on the measurement of organizational behavior and the quality of work life. New York: John Wiley and Sons, 1980.

Struening, E. L., & Guttentag, M. Handbook of evaluation research (Vol 1). Beverly Hills, CA: Sage Publications, 1975.

Suchman, E. Evaluative research. New York: Russell Sage Foundation, 1967.

Terborg, J. A., Howard, G. S., & Maxwell, S. E. Evaluating planned organizational change: A method for assessing alpha, beta, and gamma change. Academy of Management Review, 1980, 5, 109-121.

Terpstra, D. E. The OD evaluation process: Some problems and proposals. Human Resource Management, 1981, 20, 24-29.(a)

Terpstra, D. E. Relationship between methodological rigor and reported outcomes in organization development evaluation research. Journal of Applied Psychology, 1981, 66, 541-543.(b)

Terpstra, D. E. Evaluating selected OD interventions: The state of the art. Group and organization studies, 1982, 7, 402-417

Tryon, W. W. A simplified time-series analysis for evaluating treatment interventions. Journal of Applied Behavioral Analysis, 1982, 15, 423-429.

Tuchfeld, B. S. Some problems in assessing change. In L. E. Datta and R. Perloff (Eds.), Improving evaluations. Beverly Hills, CA: Sage Publications, 1979.

Umstot, D. D. Organization development strategies in the U. S. military. Group and Organization Studies, 1979, 4, 135-142.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Weisbord, M. R. Some reflections on OD's identity crisis. Group and Organization Studies, 1981, 6, 161-175.

Weiss, C. H. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1972.

White, S. E., & Mitchell, T. R. Organization development: A review of research content and research design. Academy of Management Review, 1976, 1, 57-73.

Zmud, R. W., & Armenakis, A. A. Understanding the measurement of change. Academy of Management Review, 1978, 3, 661-669.